引用格式:程聪, 陈佳晨, 严璐璐. 基于形式理性与实质理性的大模型价值对齐机制[J]. 技术经济, 2025, 44(1): 28-39.
Cheng Cong, Chen Jiachen, Yan Lulu. Research on value alignment mechanism of large language models based on formal and substantive rationality[J]. Journal of Technology Economics, 2025, 44(1): 28-39.

基于形式理性与实质理性的大模型价值对齐机制

程 聪[1,2], 陈佳晨[1], 严璐璐[1]
(1. 浙江工业大学管理学院, 杭州 310023; 2. 浙江工业大学中国中小企业研究院, 杭州 310023)

摘 要: 大模型价值对齐是关涉企业乃至社会在采用大模型技术进行安全协作的全球性议题。如何实现大模型行为与决策者的价值意图及社会规范系统保持一致,成为确保大模型应用安全性和信任度的核心问题。首先,本文引入马克斯·韦伯提出的形式理性和实质理性两个重要哲学概念,探讨大模型价值对齐机制。研究发现,大模型应用于企业管理存在四种价值对齐状态:"高形式理性-低实质理性"的技术偏移、"高实质理性-低形式理性"的价值优先、"低形式理性-低实质理性"的对齐失效及"高形式理性-高实质理性"的动态对齐。其次,甄别了透明性、清晰性和社会性三种价值对齐的分析标准。最后,构建了大模型应用于企业管理的价值对齐实现路径,包括"技术偏移→动态对齐"的认知能力具身性路径、"价值优先→动态对齐"技术意向性的清晰化路径,以及"对齐失效→动态对齐"意义建构路径。研究成果为大模型应用于企业管理的价值对齐机制提供理论支撑与实践启示。

关键词: 大模型技术; 价值对齐; 形式理性; 实质理性
中图分类号: F272.3 **文献标志码**: A **文章编号**: 1002-980X(2025)01-0028-12
DOI: 10.12404/j.issn.1002-980X.J24081809

一、引言

得益于泛化性、通用性、生成性等超人工智能潜能的持续涌现,大模型技术作为新一代人工智能的重要发展方向,成为打造新质生产力的重要引擎,推动人类社会进入全新的智能经济时代。作为第四次工业革命的核心技术,人工智能被认为将在深层次上重塑生产要素的分配格局[1]。然而,在大模型掀起社会经济发展颠覆性变革的同时,大模型威胁论日渐甚嚣尘上[2]。在此背景下,如何实现大模型行为与决策者的价值意图及社会规范系统保持一致,成为确保大模型安全性和信任度的核心问题。价值对齐指的是确保人工智能系统的行为、决策和输出符合人类预期和社会价值观,防止其产生偏离或不符合社会伦理标准的行为。尤其是当人工智能接管更多人类任务的情境下,如何保证其决策符合社会公认的道德和法律框架,显得尤为重要。大模型价值对齐概念获得众多拥趸,成为全球性人工智能发展议题。例如,我国相继发布《新一代人工智能治理原则——发展负责任的人工智能》《生成式人工智能服务管理暂行办法》《全球人工智能治理倡议》,呼吁在人工智能领域践行构建人类命运共同体理念。国际社会也高度重视促进人工智能健康发展,提出诸多倡议。例如,联合国教科文组织发布《人工智能伦理问题建议书》;二十国集团提出"G20人工智能原则";欧盟委员会提出《人工智能的伦理准则》等。

现有研究从多科学角度探讨了大模型的价值对齐问题,计算科学领域提出了基于人类反馈的强化学习、奖励模型的解决思路[3-4]。社会科学领域则从人类道德和伦理原则切入,探讨价值对齐问题的有效路

收稿日期: 2024-08-18
基金项目: 国家自然科学基金专项项目"大模型技术应用于企业战略管理优化机制研究"(72342029);国家自然科学基金重大项目"创新驱动的企业国际创业理论与战略研究"(72091314)
作者简介: 程聪(1985—),博士,浙江工业大学管理学院副院长,中国中小企业研究院研究员,博士研究生导师,研究方向:数字创新战略,企业国际化战略;陈佳晨(1998—),浙江工业大学管理学院硕士研究生,研究方向:企业创新管理;(通信作者)严璐璐(1997—),浙江工业大学管理学院博士研究生,研究方向:数字创新战略,企业国际化战略。

径。例如，Weidinger 等[5]基于“无知之幕”的哲学概念探讨大模型决策系统的公平性。Bauer 和 Vivtuous[6]则探讨了人类伦理规范下的大模型系统构建。尽管现有研究为大模型价值对齐问题解决提供了丰富的见解，但其隐含假设是确保机器目标与人类社会规范系统达成一致，而忽略了大模型决策行为满足人类任务需求或意图的另一层含义。换言之，大模型的价值对齐不仅要确保其决策行为符合社会规范，还要关注是否满足人类的实际任务需求和意图，形成兼顾伦理和效用的双重需求。

为进一步理解大模型价值对齐的双重含义，本文引入两个重要的哲学概念：形式理性和实质理性。形式理性强调方法和程序的逻辑性、计算性和规则性，关注行动的方法和手段是否符合逻辑、是否可以被计算或量化[7]。对于大模型而言，形式理性体现于系统在严格的规则和算法框架内执行任务，追求效率和精确性，以实现最佳的计算和逻辑一致性。在此框架下，大模型的设计重点在于完成任务的准确度和可量化的优化结果。相反，实质理性则是通过一组潜在的价值假设来指导行为，聚焦于行动是否合乎经验、道德和美学等方面的内容[7-8]。换言之，实质理性不仅关注大模型任务的正确执行，还强调大模型决策过程中的道德、伦理和社会责任，要求大模型在处理复杂情境时，考虑更广泛的社会影响，确保其行为符合人类社会的道德和价值规范，而不仅仅是任务导向的有效性。在此基础上，本文构建了大模型价值对齐的评估框架，甄别了大模型价值对齐的四种状态、分析标准及实现路径，为大模型价值对齐提供理论支撑与实践启示。

二、文献回顾

（一）大模型在企业管理中的应用

大模型是指在广泛且大规模数据集上通过自我监督或半监督学习方法进行预训练的具有数百万至数千亿参数的大规模预训练模型。通过在预训练过程中深入学习数据中蕴含的复杂特征和结构，而后以微调的方式快速而有效地适应多种不同的下游任务[9-10]。

大模型的功能包括模型预训练、适配微调、高效计算、推理加速等。首先，模型预训练取决于高效的策略、高质量的数据及高效的模型架构三个方面[11]。预训练策略采用优化任务目标、热启动策略等方法，以较低成本实现预训练。预训练数据的构建过程，需经过质量筛选、冗余数据去除和隐私保护，保障数据质量。预训练模型架构包括了统一的序列建模[12]，将多个自然语言处理任务整合到一个框架中，以提升模型性能和泛化能力[13]。其次，适配微调根据特定任务需求，涉及指定微调和参数高效微调[14]。指令微调使大模型根据给定的指令提示提供特定回应，有助于大模型获得人类语言指令的理解、数据获取和对齐等能力。参数高效微调涵盖在元模型的基础上微调引入额外参数、微调元模型的部分参数及将模型参数重参数化到低维度参数空间，优化低维空间中的近似参数[15]。再次，模型高效计算通过优化计算资源，提升训练吞吐量，从而在有限资源下最大化模型的高效计算[16]。最后，模型推理加速主要采用模型压缩技术[17-18]，包括模型稀疏化、模型蒸馏、模型参数共享等。

近年来，大模型在文本生成、人机交互、代码生成及基于常识或领域知识的推理方面展现出卓越的能力[19-20]，促进了其在企业管理的深入应用。大模型不仅优化了现有业务流程，还开创了智能代理构建的新范式，通过感知环境、存储和检索记忆，以及深入的反思和计划，极大地辅助了企业在制定战略决策方面的能力[20]。具体来说，大模型在企业管理中的应用具有双重意义。一方面，企业开发或集成大模型技术，旨在通过技术创新获得竞争优势，具体而言，大模型已经成为企业商业竞争的关键核心所在。例如，微软将 ChatGPT 集成至 Bing 搜索引擎，显著提升了其在在线搜索市场的竞争力，进而推动其他如谷歌、百度等技术巨头加速发展自身的大模型技术。另一方面，企业通过集成大模型来提高其业务操作的效率和效果。例如，大模型通过自然语言处理、知识图谱等技术，加速企业新药研发进程，实现高效、创新、个性化的药物设计和发现。然而，大模型在企业管理应用中暴露出两方面问题：一方面，大模型因其庞大的参数规模而带来的训练难度大、成本高，以及对文本数据和计算资源的高需求，有时难以证明其增加的实施成本的合理性[21]。另一方面，大模型的“黑箱”性质[22-23]使得外部用户难以理解模型如何将输入转化为输出，或确定各输入在输出中的相对重要性。这种不透明性可能导致大模型在执行复杂任务时出现与人类价值观不相匹配或与人类意图不一致的行为，从而增加潜在风险。

大模型应用于企业管理是一个处于不断发展中的新兴现象,针对大模型在企业管理等实践应用的研究仍相对有限[24]。目前研究主要聚焦于利用大模型、机器学习等先进研究方法展开分析,包括大模型在执行特定任务时相较于人类的准确性[25],GPT 在文本分析中的应用[26]、利用大语言模型优化数字化转型测度[27]等。然而,关于大模型在企业管理中的价值对齐问题,即确保技术输出与企业的战略目标和核心价值观相匹配的重要性,研究关注相对有限。

(二)价值对齐:大模型技术应用的社会基准

大模型技术与企业决策者共同执行复杂任务的前提,是确保大模型输出结果与人类预期目标相一致,即实现价值对齐。Wiener[28]早在1960年提出警示,自动化技术可能发展到使人类难以在决策过程中及时有效地干预。因此,确认机器学习输入的目标反映人类真实意图的步骤十分关键。近年来,学术界和产业界对价值对齐的概念进行了深入的探讨,但仍然存在模糊和分歧。例如,Russell[29]认为价值对齐涉及确保大模型的行为符合社会规范并保障其安全可靠。Gabriel[30]将价值对齐问题划分为规范挑战和技术挑战,前者关注如何使大模型行为符合社会规范,后者则探讨如何选择并引导大模型行为的价值规范。事实上,价值对齐由“价值”和“对齐”两个词构成,维基百科将价值定义为客体对于主体表现出的正面意义和有用性。因此,价值对齐概念包含两层含义,一是大模型的行为与企业决策者目标意图的一致性,二是大模型的行为与企业决策者所遵循的社会规范系统的一致性。

价值对齐的紧迫性正随着生成式人工智能的加速迭代而愈渐凸显[31]。在企业管理中,没有价值对齐的大模型不仅可能损害企业的利益,更有可能触发一系列连锁反应,最终导致企业的声誉乃至整个行业的崩盘。例如,大模型在不具备适当的价值对齐机制的情况下,可能会无意中反映并放大其训练数据中存在的性别、种族偏见或文化刻板印象[32]。不仅损害企业的多元化和包容性努力,也可能触发利益相关者的强烈反弹。此外,大模型可能会生成攻击性、毒害性言论[33]。例如,Li 等[34]发现部分大模型表现出的毒害性言论根植于心理行为,在句子分析水平上是无法捕捉的。放置于企业管理情境中,若大模型出现攻击性、毒害性言论,必定会引发在社交媒体的热议,进而引发公关危机,对企业形象和客户信任造成长期的负面影响。更为严重的是,大模型在信息处理和生成过程中可能产生“幻觉”效应,即生成与事实不符的虚假信息[35],这种信息的流布可能会误导决策者,形成错误的战略决策,从而使企业陷入严重的商业危机。

现有研究从多学科角度探讨大模型的价值对齐问题解决。在计算科学领域,研究者采用多种方法评估和提升价值对齐的质量和效率。例如,通过识别和处理关键状态,展示机器人在关键状态下的行为来建立用户的适当信任,从而解决人机交互中的价值对齐问题[36]。Christiano 等[4]通过交互和主动学习的智能体性能的渐近一致性,从而学习出一个奖励函数,训练深度强化学习模型以实现价值对齐。Brwn 等[3]则通过定义价值对齐和构建最小查询测试来进行精确验证、近似验证和启发式验证,构建了价值对齐验证的理论分析框架。在社会科学领域,研究者尝试将人类道德和伦理原则融入到大模型决策过程中。例如,Awad 等[37]提出利用人类的道德直觉作为指导,开发出与人类道德保持一致的机器学习模式,有效解决价值对齐问题。Weidinger 等[5]探讨如何利用“无知之幕”这一哲学概念来设计公平的大模型系统,使之在不知道未来将服务于哪类人群的情况下做出决策,实现广泛的社会公正。Bauer[6]则从哲学和伦理学视角触发,探讨如何构建能够遵守人类伦理规范的智能系统,确保它们的行为不会偏离人类的伦理和道德标准。尽管上述研究为大模型价值对齐问题解决提供了丰富的见解,但其隐含假设是确保机器目标与人类价值观的一致性,而忽略了大模型的行为满足企业决策者具体任务需求,即意图一致性。因此,亟需探讨如何共同确保大模型行为与企业决策者的意图一致性和价值一致性。

三、大模型价值对齐的评估框架构建

(一)大模型的价值对齐评估框架

1. 价值对齐的理论基础

随着人工智能技术的迅速发展,特别是大模型技术的广泛应用,更具挑战性的问题随之而来:如何实现价值对齐?本文借鉴马克斯·韦伯提出的形式理性和实质理性作为价值对齐的理论基础。形式理性涉及

方法和程序的逻辑性、计算性和规则化，关注行动的方法和手段是否符合逻辑、是否可以被计算或量化。形式理性将可计算性的最大化视为核心目标[7,38]。例如，Lindebaum 等[8]指出，基于人工智能的算法是形式理性的超级载体，它通过逻辑和数学程序操作以优化目标。一方面，人工智能在执行任务时进行“学习”，在给定约束或边界条件下，在数据集中找到逻辑上或数学上“正确”的解决方案。另一方面，人工智能因处理大规模数据并以前所未有的速度产生或计算结果成为超级载体，尽管这些方式通常难以追踪甚至是不透明的。实质理性则是通过一组潜在的价值假设来指导行为，聚焦于行动是否合乎经验、道德和美学等方面的内容[7-8]。

形式理性和实质理性之间存在复杂且动态的关系[8,40]。一是理性形式的主导地位差异。当一种理性形式强加于另一种理性形式之上时，可能会失去对技术的掌控。正如科幻小说《The Machine Stops》中所描述：“我们创造了机器，让它服从我们的意志，但现在我们无法让它服从我们的意志了。机器在发展，但不是按照我们的路线。机器在前进，但不是朝着我们的目标。”[36-37,39]二是理性形式的动态转化。实质理性可能包括基于特定文化或经验的具体道德判断或标准，当这些判断或标准被抽象化为一套普遍原则时，将转化为形式理性。届时，其不再依赖于具体的情境细节，而是被普遍地应用到各种差异化的情境中。例如，谢小云等[40]指出，数字化技术往往缺乏实质理性，甚至会把所有关乎道德、审美、目的和价值的实质理性问题转化为形式理性问题，通过计算得失风险的方式来进行价值选择，从而可能惩罚任何偏离计算得出最优解的行为。

2. 价值对齐的评估框架

形式理性、实质理性及其两者复杂关系的探讨为构建大模型应用于企业管理的价值对齐评估提供了理论基础。在此基础上，本文构建了价值对齐的评估框架，如图 1 所示。

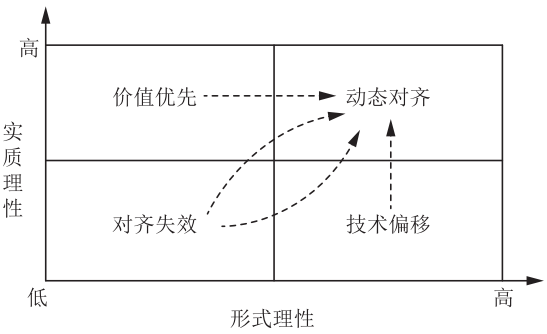

图 1　价值对齐的评估框架

（1）“高形式理性-低实质理性”的技术偏移。当形式理性高而实质理性低时，大模型应用于企业管理的价值对齐状态可能表现为以一种逻辑严密、结构完整的方式满足企业战略决策者的目标的表层意图。但对其决策可能产生的实质性、伦理性影响则视而不见。大模型技术通过计算“更正确”的解决方案来学习和改进[8]。“正确性”是基于历史数据的预测准确性而非判断和理解。因此，Balasubramanian 等[41]提出大模型技术在本质上并不是进行感知学习。相反，大模型技术主要依赖于嵌入在统计模型的形式理性，用以优化对输入数据的预测效果。亚马逊的自动化绩效管理系统生动地体现了“高形式理性-低实质理性”的技术偏移。该系统严格依赖数据，实时监控员工的工作效率，记录任务完成速度、休息时间等指标，对未达标员工自动发出警告，甚至可能直接触发解雇决策。《华盛顿邮报》《纽约时报》等媒体多次揭示了该系统对员工身心的潜在影响，指出员工在高压、高强度的管理模式下，普遍面临着巨大的心理和身体负担。这种系统显然突出了自动化技术在效率与准确性上的优势，被外界广泛认为过于机械化，缺乏人文关怀和社会责任的考虑，反映出高形式理性在追求效率最大化的过程中对实质性关怀的疏忽。正如 Bostrom[42]在《Superintelligence：Paths，dangers，strategies》著作中提到的“回形针最大化”思想实验也反映了这种技术偏移：想象人类给一个超级智能的人工智能系统设定优化回形针生产的目标，人工智能系统在缺乏对社会、环境、全局的思考下，可能会有效动用所有资源转化为回形针。即使人类的初衷并不是打算毁灭人类来制造更多的回形针，但因其在提示词中忽略了这一点。此外，大模型决策根据提示词的制定方式，依赖于形式理性或非个性化的定量计算，快速产生特定回答，大幅度降低了决策的内在多样性和复杂性[41]。此类技术决定论的倾向[43]忽视了创新过程中的多元性和伦理复杂性，导致大模型决策机制缺乏透明度和可解释性。

（2）“高实质理性-低形式理性”的价值优先。在实质理性占优而形式理性较弱的情境下，大模型在企业管理中的应用表现出对社会伦理价值的高度敏感性，但相对弱化了对效率最优的决策逻辑。此时，大模型应用虽然符合社会规范系统，但在决策的执行一致性和效率上可能有所欠缺。阿里巴巴的“绿色物流”项目

是“高实质理性-低形式理性”价值优先的典型案例。阿里巴巴集团旗下“菜鸟网络”物流平台利用大模型分析历史数据、供应链模式和用户需求,推荐更环保的包装方案,帮助减少一次性塑料和不可降解材料的使用。同时,大模型根据用户分布和运输需求,动态优化运输路线,以实现燃料消耗和二氧化碳排放的双重降低,推动“绿色低碳”物流的落地。这一大模型应用超越了单纯追求效率最大化的目标,而是旨在支持阿里巴巴的“碳中和”承诺,进一步彰显其履行社会责任的决心。然而,这种非形式化的决策模式也引发了争议。部分学者认为,过度依赖大模型中嵌入的非结构化价值判断,可能导致输出结果在执行上缺乏一致性,从而增加组织决策中的不确定性与潜在不信任[44]。与此同时,管理者在实际应用中通过实质理性来反思和调适这些决策,使其更符合动态变化的价值需求[8]。因此,大模型的这种应用虽然不完全遵循效率导向,但在人类固有的创造力和多元价值观的作用下,能够丰富企业的决策多样性和情境适应性,为企业在复杂环境中提供了更具包容性和社会价值导向的思维方式[41]。

(3)“低形式理性-低实质理性”的对齐失效。当形式理性和实质理性均处于较低水平时,大模型在企业管理决策中的应用既缺乏严密的逻辑和结构化过程,也未能遵循社会伦理价值。这种模式往往会导致无效甚至有害的决策实施,危及企业持续发展和社会声誉。在当今迅猛发展的大模型时代,不少在位企业在追求创新和竞争优势的过程中,倾向于快速采用新兴的大模型技术,常在未充分考虑其适用性或潜在后果的情况下将其引入关键业务决策。例如,亚马逊曾开发过一款大模型驱动的自动化招聘工具,旨在优化招聘流程。然而,这一工具在实际应用中表现出明显的“低形式理性-低实质理性”特征,最终导致对齐失效。具体而言,该系统在筛选简历时缺乏透明和严密的逻辑,无法保证人才筛选的客观性与一致性。由于训练数据源自亚马逊过往十年内的招聘记录,而男性在这些记录中比例较高,该系统在实际应用中表现出明显的性别偏见,倾向于排除女性候选人。这一偏见不仅违背了公平招聘的伦理要求,也缺乏对筛选过程的结构化控制,未能在形式上严格保障公平性。正因为大模型在逻辑严谨性和伦理价值上的双重不足,亚马逊在此系统遭到公众质疑后最终选择放弃。价值失效的原因在于两方面。一方面,企业可能将技术的先进性置于其适用性或道德后果之上。尽管大模型因其卓越的数据处理能力而备受推崇,但缺乏对这些工具如何深刻影响决策过程的理解可能会导致滥用或误用。例如,如果未能适当地配置和监管,大模型可能在未经验证的情况下自动化重要决策,忽略潜在的错误或偏见,从而产生不利的商业和社会结果。另一方面,在没有建立有效的伦理监管机制的情况下,使用大模型的决策过程可能导致侵犯隐私、数据滥用和其他伦理问题,这不仅可能损害消费者信任,还可能触发法律和道德争议。

(4)“高形式理性-高实质理性”的动态对齐。当形式理性和实质理性均处于较高水平时,大模型应用于企业管理达到了价值对齐的最优状态——动态对齐,不仅决策过程合乎逻辑、规范,决策结果也具有高度的道德和实用价值,能够有效促进组织和社会的整体福祉。腾讯医疗 AI 项目“腾讯觅影”是大模型价值动态对齐的典型案例。通过大模型驱动的图像识别技术,“腾讯觅影”能够快速、精准地分析医学影像,辅助医生在早期筛查和诊断中做出准确判断,显著提高诊断效率。尤其是在癌症筛查、肺结核检测等领域,这项技术帮助医生在短时间内处理大量病例,减少了误诊和漏诊的可能。不仅如此,“腾讯觅影”还将这一技术应用到偏远地区,为当地基层医疗机构提供实时的辅助诊断支持。通过远程影像分析和诊断,偏远地区的医生可以借助这一技术获得专业支持,帮助当地患者在初级诊疗阶段获得及时、有效的医疗建议,彰显了形式理性与实质理性之间取得平衡能够有效促进社会福祉。现有研究也指出,组织乃至人类社会实际上构成了一个异质的价值体系,其中不同的组织单元表现出不同的社会价值需求[45]。这引出了一个核心问题:当大模型深入整合到企业管理实践中,能否灵活地根据差异化的决策者价值诉求实现动态对齐?本文认为,通过有效运用主导逻辑,即企业高层管理团队用于指导决策和解释环境信息的知识结构与认知框架[46],不仅塑造了组织如何识别和响应外部环境的方式,而且还定义了其内部价值观的表达和实践。在这一框架下,大模型可以被视为加强和优化这一逻辑的工具,使其更加精确地匹配组织的长期目标与市场动态。进一步,为了应对组织内部和外部的多样化价值需求,大模型需要设计成可适应性强的系统,能够捕捉和分析来自不同利益相关者的反馈,从而支持形成广泛共识的战略决策。例如,通过数据驱动的洞察力,大模型可以帮助企业管理层平衡多元化的利益诉求,实现更为公平的价值对齐。

（二）大模型的价值对齐实现路径

1. 价值对齐的分析标准

如何构建一个具有操作性的价值对齐实现路径尤为迫切，本文提出价值对齐的三个分析标准。

第一，透明性。透明性指人类对大模型技术应用于企业管理的掌握能力，即实现大模型技术应用与人类认知能力之间的平衡性。大模型应用于企业管理的前提是确保所有利益相关者都能对大模型的功能及其局限性、操作原理，以及如何有效地操控大模型输出拥有全面而深入的理解。《中国人工智能系列白皮书——大模型技术（2023 版）》指出，大模型的不透明性增加了定位和解决任务错误的难度，因此透明性的确保是技术开发的基本要求，更是道德和法律责任的体现。透明性涵盖了对大模型决策过程的透彻洞察，以及对其在执行各项任务时的表现和行为的详尽分析。尽管像其他先进技术的设计过程类似，人类参与了大模型的训练和测试过程，但大模型存在诸多显著差异：相较于依据人类规定理解和执行任务的其他先进技术，大模型不局限于任何特定的本体论故事，也不依赖于任何离散的、基于对象的形式化本体论[47]。相反，基于统计分析，大模型技术从历史数据中自主地推断潜在的决策规则，而这些推断的规则通常是不透明的，即便是大模型技术的开发者也无法直接观察或操纵规则，只能看到最终的输出结果。因此，大模型技术的不透明性增加了定位和解决任务错误的难度[41]。此类错误的识别通常还存在滞后性，只有在错误发生并可能影响到诸多决策之后，才会被人类决策者所注意，因为人类并非对大模型技术作出的每一个决策进行过滤[48]。因此，透明性的确保不仅仅是技术开发的基本要求，更是道德和法律责任的体现。

第二，清晰性。清晰性作为大模型在企业管理中实现价值对齐的关键指导原则，其核心在于确保决策过程中价值观的表达不仅精确无误，而且能够全面地反映人类的复杂价值体系。一方面，清晰性要求企业在部署大模型之前，必须明确其价值对齐的目标，并且确保这些目标经过精心设计和表述，以清晰地反映企业的核心信念和道德标准。这意味着企业需要在技术部署前进行全面的利益相关者分析，以识别和整合利益相关者期望，并将其明确地嵌入至大模型决策框架中。另一方面，清晰性不仅要求技术实现上的精确性，更要求大模型能够在不同的应用场景中灵活地适应企业价值观的演变。例如，采用先进的深度学习等机器学习技术，通过连续学习优化决策模型，确保大模型随着时间的推移和环境的变化而逐渐适应企业战略目标的演变。

第三，社会性。社会性是指大模型决策结果符合人类社会规范系统，同时人类赋予大模型决策结果以现实意义，是大模型应用于企业管理的客观需要。一方面，确保大模型在企业管理应用和发展过程中始终服务于人类，并且其行为和决策遵循人类的价值观和伦理标准。这将保证即便未来大模型发展成拥有媲美甚至超越人类能力、拥有自我意识的超人工智能，其嵌入的道德原则、伦理规范和价值观，也必须与人类的道德原则、伦理规范和价值观保持一致。另一方面，企业决策者通过了解大模型技术如何运作来回应技术解决企业管理问题的现实意义。因此，只有在找到关于大模型技术在解决实际问题过程中的合理技术解释时，企业决策者才能接受大模型技术输出结果。换言之，大模型技术的价值兑现取决于其分析结果是否被决策者所采信。一是，决策者需要保持对大模型技术应用过程中的社会意义建构的热情。大模型技术本意虽然是帮助决策者减轻认知负担，但同时也会弱化决策者对于大模型技术应用的解释欲望，聚焦于大模型技术可行性，而相对忽略技术分析结果的社会意义建设[49]。从本体论观点来看，过渡关注大模型技术“事实”，而忽略技术背后的社会解释性价值，只能产生技术与社会相脱节，问题与结果两张皮的现象。二是，大模型技术应用的社会意义建构是一个系统性的全局过程。这意味着弱化大模型技术分析的每个步骤等细节性解释，更加关注整个技术运用产生结果的社会价值。当然，并不是说大模型技术运用整体性的社会意义建构离开技术分析细节而独立存在，而是如何在大模型技术分析具体环节与社会整体性上实现协调性平衡，这也是大模型决策面临的又一重要议题。

2. 价值对齐的实现路径

上述讨论可知，“动态对齐”是大模型应用的理想状态，确保决策过程逻辑严谨、符合规范，且具有高度的道德和实用价值。但在特定情境下，“价值优先”状态同样能够在社会责任和伦理价值上发挥积极作用，

为社会福祉做出贡献。在路径实现分析中，本文只考虑动态演变的“起点”和“终点”，至于“中间状态”的多样性有待未来研究进一步深化。大模型应用的价值对齐框架图如图 2 所示。

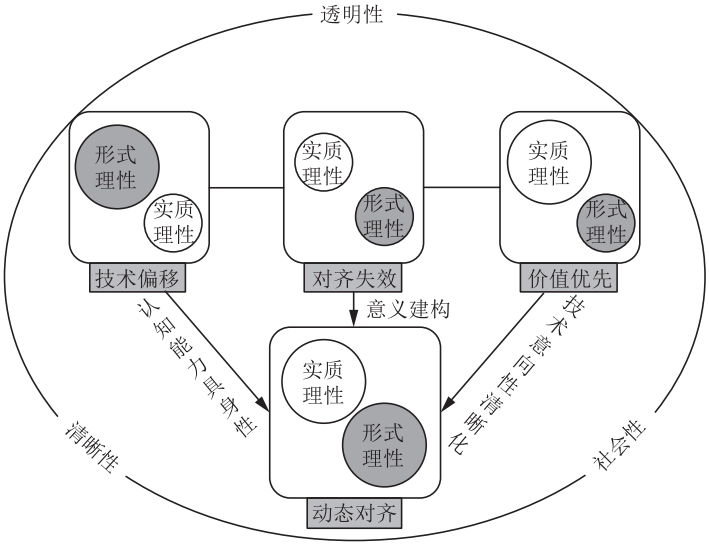

图 2　大模型应用的价值对齐框架

（1）“技术偏移→动态对齐”的认知能力具身性路径。大模型在企业管理中的应用本质上诉诸于对人类认知活动的模拟范围或类型不断扩展，从而不断提升大模型应用的认知能力[50]。因此，大模型技术借助人类固有的本能认知、经验认知和推算认知对企业管理情境进行自适应、自学习和自组织[50]，从而逐渐夯实价值对齐的底层伦理逻辑，推动“技术偏移”向“动态对齐”转变。具体而言，本能认知是个体基于本性对社会环境形成反应的认知活动，如直觉。经验认知是个体在行动和学习过程中对社会环境进行反应的认知活动，如直接经验和间接经验。推算认知则是个体基于后天学习所具备的推理、计算能力的认知活动，所获得的知识通常是理性的产物[50]。面对企业管理乃至人类社会经济系统中的深层次问题，大模型决策应在缓解人类认知压力的目标导向下，将数据、信息和知识先后嵌于人类本能认知、经验认知和推算认知框架，而非塑造极端复杂的符号表征困境[51]。换而言之，大模型决策系统应当作为决策管理者理解世界的媒介和“感官”，即提升大模型系统的“具身性”。同时，大模型决策系统应当模拟人类与社会环境的互动，继承社会规范系统。为此，应从认知冲突察觉、意义协商、权力重构和共情培育四个方面，彰显大模型知觉透明性和覆盖人类价值的导向。第一，认知冲突察觉。在推算认知的指导下，技术开发者不仅能够察觉与其他利益相关者之间的认知差异和重估，还能够预测和理解技术应用于企业管理可能带来的认知挑战和冲突。第二，意义协商。在认知冲突觉察的基础上，进一步推动技术偏移向动态对齐的转变需要意义协商。学习认知强调知识的习得和共享，有助于利益相关者在意义协商过程中理解和接纳彼此观点，形成一个多元包容的意义协商结果。第三，权力重构。行为认知的运用有助于理解和调节不同利益相关者之间的权力关系和话语权分配，从而实现意义建构过程的公正和平等。第四，共情培育。本能认知促使利益相关者更加关注人性化和情感因素，在共情培育过程中培养相互理解、尊重和信任，从而实现技术与价值的动态对齐。

例如，特斯拉的自动驾驶技术在数据处理和算法优化方面展现出极高的形式理性，通过深度学习技术使系统具备自我进化的能力，从而不断适应新的驾驶场景。然而，特斯拉的全自动驾驶系统在应对复杂或非预期场景时，仍可能出现误判或反应延迟的情况，进而带来潜在的交通安全隐患。这反映了特斯拉在技术创新的早期阶段，更多侧重于技术性能的提升，而对可能产生的负面影响和伦理风险考虑相对不足，反映出“技术偏移”下形式理性过强、实质理性不足的问题。而后，特斯拉逐步通过定期的软件更新和透明的事故报告机制，加大了在自动驾驶技术开发中对安全性和伦理问题的关注。随着这些改进措施的落实，特斯拉在技术创新与社会价值的对齐方面逐渐达到了较高水平，实现了从“技术偏移”到“动态对齐”的转变。这一过程展示了企业在追求技术领先的同时，如何通过提升实质理性来平衡技术创新与社会责任之间的关系。

（2）“价值优先→动态对齐”技术意向性的清晰化路径。当前，大模型在企业管理中的应用存在传统认识论意义上的不透明性。面对庞大且复杂的计算量，人类无法审查所有的计算过程，这不仅导致人类在认知上存在着盲区[52]，而且削弱了决策者对大模型输出结果的可信度。因此，实现“价值优先”向“动态对齐”转变取决于技术意向性的清晰化。在解释关系中，大模型技术作为物理世界的表征，提供抽象化和概念化的表征来抽象而真实地反映企业决策者面对的管理问题，这种反映关系被称为诠释学的透明性。虽然大模型系统的表征诠释取决于决策者既有知识结构能否对大模型表征的经验诠释，然而大模型系统对表征与世

界之间的映射关系则更加缄默地决定了表征与世界之间的意向是否被决策者正确解读，这种结构被称为意向弧。因此，当技术指向现实的意向性越清晰，相对应的诠释关系越强，大模型决策结果与决策者预期目标的一致性更高。技术意向性的清晰化路径包括人类价值偏好学习、可扩展监督、嵌入式代理等。人类价值偏好学习是利用大模型分析企业决策者行为、选择和反馈来学习和模仿决策者的价值偏好，从而帮助大模型在没有明确指令的情况下做出符合决策者期望的决策，包括数据驱动的学习、反馈循环、模拟决策场景等多种方法。可扩展监督强调监督机制的扩展性，在大模型应用的扩展和复杂性增加的同时，进行升级，包括分层监督、动态调整、多维度评估等。嵌入式代理则是将人类决策逻辑嵌入到大模型应用中，使其在独立运行时做出符合预定价值观的决策，包括策略嵌入、自主调节、持续更新等。

例如，谷歌在医疗领域推出的 Pathways 大模型系统，旨在通过分析和整合庞大的医学数据来辅助医生诊断，尤其是在影像识别和疾病预测方面提供决策支持。起初，Pathways 的设计更多聚焦于帮助医生在短时间内筛查出潜在风险。虽然系统在数据处理能力和诊断效率上表现优异，但由于计算过程的复杂性，Pathways 的决策过程不完全透明，医生难以审查模型的所有计算步骤，这种不透明性降低了对大模型输出结果的可信度。意识到这种“价值优先”状态的局限后，谷歌逐步推动 Pathways 向“动态对齐”转变，以增加系统的技术意向性和决策透明性。在这一过程中，Pathways 团队采用了人类价值偏好学习和可扩展监督等方法。通过分析医生的诊疗行为和反馈，Pathways 在没有明确指令的情况下逐步学习和模仿医生的诊疗偏好，形成数据驱动的反馈循环，确保其诊断建议符合医生的预期。与此同时，Pathways 引入了可扩展监督机制，分层对诊断建议进行监控和动态调整，利用多维度评估体系提升模型输出的可信度和透明度。Pathways 还采用了嵌入式代理，将医生的诊疗逻辑嵌入模型中，使系统在独立运行时也能做出符合医学伦理和专业标准的决策。这种意向性的清晰化路径确保了 Pathways 不仅能够高效地处理诊断任务，还能在技术表现上与医生的伦理和专业标准保持一致，从而增强了医生对大模型系统的信任度，实现了技术创新与医学伦理的高度融合。

（3）“对齐失效→动态对齐”意义建构路径。推动大模型在企业管理中的“对齐失效”向“动态对齐”转变依赖于意义建构。意义建构是通过外部信息理解、认识、预测并改造世界的过程，是个体进行决策的基础[53]，主要依赖于两类信息的联结：一是自外部信息源而来的线索，二是自信息接收者过去经验和阅历所构建的认知框架，两者的结合构建了信息意义。大模型决策中任务情境与信息线索的联结是脆弱的。一方面，受制于信息接收者的先验认知框架能否接纳与预期不符甚至相反的决策结果，另一方面，也受制于大模型系统对数据、信息和知识意义的有效收集和解读。因此，“对齐失效”向“动态对齐”转变的意义建构路径包括信息意义的构建、任务情境的嵌入、认知框架的适应及反馈循环的迭代。信息意义的构建是指通过分析和解释数据来构造有意义的信息，使决策者能够根据这些信息做出知情决策。在大模型的上下文中，这个过程涉及将复杂的数据和算法输出转换成易于理解和操作的格式。为了实现这一点，模型需要被设计为能够识别关键信息，突出显示对决策过程最关键的数据。任务情境的嵌入强调的是将决策模型放在具体的操作环境中使用，确保模型输出与实际情境紧密相关。这要求模型能够捕捉和理解其运行环境的特定条件，包括业务规则、市场动态和组织目标等。通过将这些上下文信息编码到模型中，可以提高模型的相关性和应用价值。认知框架的适应涉及调整决策者对信息处理和解释方式的认知结构，使之更好地适应由大模型提供的新信息和方法。在大模型应用中建立一个有效的反馈循环是确保持续改进和优化的关键。这包括收集决策者的使用反馈、性能数据和业务成果，然后根据这些信息调整模型参数和运行策略。迭代的反馈循环不仅可以提高模型的精确度和效率，还可以逐步调整模型的操作，以更好地满足决策者的需求和适应变化的环境。

例如，淘宝的个性化推荐系统完善反映了由“对齐失效”逐步转向“动态对齐”的路径。早期，淘宝的推荐系统主要基于用户的历史浏览和购买数据，通过简单的关联算法对用户偏好进行预测。虽然在短期内提升了点击率和转化率，但因过度依赖过往行为数据，系统未能灵活响应用户兴趣的动态变化，导致推荐内容重复，用户体验单一，甚至出现“疲劳感”。为提升推荐效果，淘宝引入了更强大的大模型算法和自然语言处理技术，对用户的实时评论、分享等行为进行分析，提取出更细微的偏好信息。例如，用户在近期对某类新

产品表现出关注,系统便能够自动捕捉并及时响应这一兴趣变化,不再仅仅依赖用户的过往行为数据。这种改进帮助推荐系统形成了更精准的用户画像,实现了信息意义的构建,使推荐内容更具个性化和时效性。淘宝还对推荐系统开发团队的认知框架进行了优化,不再只追求点击率,而是更加重视用户体验和长期关系的建立。通过对用户行为的大数据分析,团队逐步认识到用户不仅希望获得符合个人偏好的推荐,还渴望探索新的商品和品牌。为此,系统在推荐中加入了新产品和潜在兴趣点,让用户在熟悉的商品之外,能够发现更多符合个人风格的内容,从而增强用户对平台的黏性和满意度。淘宝还构建了反馈循环机制,以实现推荐系统的持续优化。

四、研究结论与启示

(一)研究结论

本文基于形式理性和实质理性张力视角探讨了大模型在企业管理应用中的价值对齐机制,得出以下结论:

第一,大模型应用于企业管理中的价值对齐存在四种可能的状态:技术偏移、价值优先、对齐失效和动态对齐。当形式理性高而实质理性低时,价值对齐状态可能表现为一种逻辑严密、结构完整的方式完成企业决策者目标的表层意图,而不论其可能产生的社会伦理层面的影响。当实质理性高而形式理性低时,大模型在企业管理中的应用表现为对社会伦理价值的高度敏感,优先满足社会责任和价值导向。然而,由于缺乏严密的逻辑和结构化框架,模型在应用中可能在操作效率和决策一致性上存在不足。此时,企业决策者更多地将大模型作为传递和落实价值观的工具,使其决策偏向价值导向而非单纯效率,关注长远的社会效益。当形式理性和实质理性处于双低水平时,大模型应用既缺乏严密的逻辑和结构化的过程,也无法符合社会伦理规范,这将导致无效甚至有效决策。当形式理性和实质理性均处于较高水平时,大模型应用于企业管理达到了动态对齐状态,不仅决策过程满足逻辑,决策结果也符合社会道德伦理规范。

第二,甄别了“透明性-清晰性-社会性”大模型应用于企业管理的价值对齐分析标准。透明性要求决策者对大模型技术应用具有掌控力,实现大模型技术应用和人类认知之间的平衡。清晰性则要求确保大模型决策过程中价值观的精准表达,同时反映人类复杂价值体系。社会性是指大模型决策结果符合人类社会规范系统,同时赋予大模型决策结果以现实意义,服务于企业管理的客观需要。

第三,构建了大模型应用于企业管理的价值对齐实现路径,包括“技术偏移→动态对齐”的认知能力具身性路径、“价值优先→动态对齐”技术意向性的清晰化路径及“对齐失效→动态对齐”意义建构路径。在认知能力具身性路径中,大模型技术不断借助人类固有的本能认知、经验认知和推算认知对企业管理情境进行自适应、自学习和自组织,扩展人类认知活动的模拟范围,提升大模型应用的认知能力。技术意向性的清晰化路径下,不断提高技术指向现实的意向性的清晰化,增强反映的诠释关系,使得大模型决策结果与决策者预期目标的一致性得以提升。在意义建构路径中,通过信息意义构建、任务情境嵌入、认知框架适应及反馈循环迭代,不断提升大模型理解、认知、预测企业管理决策的效能。

(二)实践启示

首先,企业决策者在将大模型技术应用于企业管理时,需要权衡形式理性和实质理性。依赖形式理性能够提升效率和逻辑严密性,但若忽视伦理和社会影响,可能导致“技术偏移”问题。在需要快速适应变化的环境和需求的场景中,如实时交互系统、在线推荐系统等,“动态对齐”是理想的状态,即大模型在逻辑清晰、数据支持的基础上,同时满足社会伦理价值的要求,从而服务于企业的长期战略需求。在涉及重大伦理决策的特定情境中,如医疗诊断、法律判断等领域,需要确保大模型系统的行为符合社会伦理标准,保护用户的基本权利和福祉。政策层面可鼓励推行社会责任导向的技术指南,以确保企业在创新时关注其社会影响,避免因过度追求效率而忽视社会责任。

其次,企业在应用大模型技术的过程中,需要不断增强大模型的透明度、清晰度和社会性。这意味着模型的运作原理、决策依据应当对相关利益者透明,确保企业内部及外部利益相关方对模型功能、局限性和运行机制有清晰理解。在这种情况下,“动态对齐”有助于企业在决策中保持掌控力、快速纠偏,并加强与利益

相关者之间的信任。企业可以将技术逻辑与社会、文化和伦理因素结合，通过多元化数据输入与价值整合，使模型输出更具包容性和社会适应性。监管机构可推行大模型透明合规标准，指导企业进行透明披露，包括对决策依据、模型逻辑和关键变量的说明，确保公众和监管方对模型的透明理解，从而提升信任度并赋予企业动态调整的灵活性。

最后，大模型训练过程中引入多元化的数据源和多维度的价值评估标准。企业在部署大模型时，既要确保其符合当下的社会规范和伦理标准，又要灵活应对外部环境的变化，服务于企业的实际战略需求。在商业和社会责任之间找到动态平衡时，“动态对齐”尤为重要。通过将大模型决策与企业长期战略目标深度整合，企业不仅可以提升自身的竞争力，还能够塑造出符合社会预期的责任导向形象，真正实现“高形式理性-高实质理性”的动态对齐。

（三）研究局限与展望

本文也存在以下两点不足，需要未来进一步研究。

第一，本文采用二分法区分形式理性和实质理性的高低状态，以简化模型和分析过程，便于理解和操作，但这一方式忽略了理性表现的连续性和复杂性。实际上，形式理性和实质理性并非截然对立，而是可以表现为各自的连续体。未来研究可以超越二元对立视角，将形式理性和实质理性视为动态变量，进一步探索两者在不同水平下的复杂交互的动态关系，及其对价值对齐状态的动态影响，进而揭示大模型在企业管理中更丰富的应用模式和潜在影响。

第二，本文在探讨大模型技术在企业管理中的价值对齐路径时，主要依赖理论分析，缺乏具体的实践案例的详尽数据支持。由于大模型技术在企业管理中的应用还处于初期阶段，实际应用案例较少，这限制了对其具体效果和潜在问题的全面理解。未来研究可以通过收集更多企业的实践案例，尤其是结合实际应用的数据和反馈，深入探讨大模型在不同情境中的价值对齐路径，揭示在实践中可能存在的中间路径和多样化的转变过程，以更加全面、实证的方式完善大模型应用的理论模型。

参考文献

[1] 黄旭. 人工智能的三种效应：理论分析[J]. 技术经济，2022，41(7)：83-92.

[2] 梅亮，陈劲，吴欣桐. 责任式创新范式下的新兴技术创新治理解析——以人工智能为例[J]. 技术经济，2018，37(1)：1-7.

[3] BROWN D S，SCHNEIDER J，DRAGAN A，et al. Value alignment verification[M/OL]. arXiv，2020[2024-07-18]. https：//arxiv. org/abs/2012. 01557. DOI：10. 48550/arXiv. 2012. 01557.

[4] CHRISTIANO P F，LEIKE J，BROWN T B，et al. Deep reinforcement learning from human preferences[M/OL]. arXiv，2017. [2024-07-18]. https：//arxiv. org/abs/1706. 03741. DOI：10. 48550/arXiv. 1706. 03741.

[5] WEIDINGER L，MCKEE K R，EVERETT R，et al. Using the veil of ignorance to align AI systems with principles of justice[J]. Proceedings of the National Academy of Sciences，2023，120(18)：e2213709120.

[6] BAUER W A. VIRTUOUS V S. Utilitarian artificial moral agents[J]. AI & SOCIETY，2020，35(1)：263-271.

[7] WEBER M. Economy and society：An outline of interpretive sociology[M]. Berkeley：University of California Press，1978.

[8] LINDEBAUM D，VESA M，DEN HOND F. Insights from “the machine stops” to better understand rational assumptions in algorithmic decision making and its implications for organizations[J]. Academy of Management Review，2020，45(1)：247-263.

[9] BOMMASANI R，HUDSON D A，ADELI E，et al. On the opportunities and risks of foundation models[M/OL]. arXiv，2022[2024-07-18]. https：//arxiv. org/abs/2108. 07258. DOI：10. 48550/arXiv. 2108. 07258.

[10] HUANG A H，WANG H，YANG Y. Finbert：A large language model for extracting information from financial text[J]. Contemporary Accounting Research，2023，40(2)：806-841.

[11] TURC I，CHANG M-W，LEE K，TOUTANOVA K. Well-read students learn better：On the importance of pre-training compact models[M/OL]. arXiv，2019[2024-07-18]. https：//arxiv. org/abs/1908. 08962. DOI：10. 48550/arXiv. 1908. 08962.

[12] RAFFEL C，SHAZEER N，ROBERTS A，et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of Machine Learning Research，2020，21(140)：1-67.

[13] HU Z，DONG Y，WANG K，et al. GPT-GNN：Generative pre-training of graph neural networks [M/OL]. arXiv，2020[2024-07-18]. https：//arxiv. org/abs/2006. 15437. DOI：10. 48550/arXiv. 2006. 15437.

[14] OUYANG L，WU J，JIANG X，et al. Training language models to follow instructions with human feedback[J]. Advances in Neural Information Processing Systems，2022，35：27730-27744.

[15] DING N，QIN Y，YANG G，et al. Delta tuning：A comprehensive study of parameter efficient methods for pre-trained language models[M/OL].

arXiv, 2022[2024-07-18]. https://arxiv.org/abs/2203.06904. DOI: 10.48550/arXiv.2203.06904.
[16] RAJBHANDARI S, RUWASE O, RASLEY J, et al. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning[M/OL]. arXiv, 2021[2024-07-18]. https://arxiv.org/abs/2104.07857. DOI: 10.48550/arXiv.2104.07857.
[17] LIANG T, GLOSSNER J, WANG L, et al. Pruning and quantization for deep neural network acceleration: A survey[J]. Neurocomputing, 2021, 461: 370-403.
[18] CHEN Y, ZHENG B, ZHANG Z, et al. Deep learning on mobile and embedded devices: State-of-the-art, challenges, and future directions[J]. ACM Computing Surveys, 2020, 53(4): 1-37.
[19] GHAFFARZADEGAN N, MAJUMDAR A, WILLIAMS R, et al. Generative agent-based modeling: An introduction and tutorial[J]. System Dynamics Review, 2024, 40(1): e1761.
[20] SUN Y, ZHANG Q, BAO J, et al. Empowering digital twins with large language models for global temporal feature learning[J]. Journal of Manufacturing Systems, 2024, 74: 83-99.
[21] STRUBELL E, GANESH A, MCCALLUM A. Energy and policy considerations for deep learning in NLP[M/OL]. arXiv, 2019[2024-07-18]. https://arxiv.org/abs/1906.02243. DOI: 10.48550/arXiv.1906.02243.
[22] CASTELVECCHI D. Can we open the black box of AI?[J]. Nature News, 2016, 538(7623): 20.
[23] LAURIOLA I, LAVELLI A, AIOLLI F. An introduction to deep learning in natural language processing: Models, techniques, and tools[J]. Neurocomputing, 2022, 470: 443-456.
[24] CORNELISSEN J, HÖLLERER M A, BOXENBAUM E, et al. Large language models and the future of organization theory[J]. Organization Theory, 2024, 5(1): 1-15.
[25] VAN VEEN D, VAN UDEN C, BLANKEMEIER L, et al. Adapted large language models can outperform medical experts in clinical text summarization[J]. Nature Medicine, 2024, 30(4): 1134-1142.
[26] 李春涛, 闫续文, 张学人. GPT 在文本分析中的应用: 一个基于 Stata 的集成命令用法介绍[J]. 数量经济技术经济研究, 2024, 41(5): 197-216.
[27] 金星晔, 左从江, 方明月, 等. 企业数字化转型的测度难题: 基于大语言模型的新方法与新发现[J]. 经济研究, 2024, 59(3): 34-53.
[28] WIENER N. Some moral and technical consequences of automation: As machines learn they may develop unforeseen strategies at rates that baffle their programmers[J]. Science, 1960, 131(3410): 1355-1358.
[29] RUSSELL S. Human compatible: Artificial intelligence and the problem of control[M]. New York: Viking, 2019.
[30] GABRIEL I. Artificial intelligence, values, and alignment[J]. Minds and Machines, 2020, 30(3): 411-437.
[31] 胡正荣, 闫佳琦. 生成式人工智能的价值对齐比较研究——基于 2012—2023 年十大国际新闻生成评论的实验[J]. 新闻大学, 2024(3): 1-17.
[32] ABID A, FAROOQI M, ZOU J. Persistent anti-Muslim bias in large language models[M/OL]. arXiv, 2021[2024-07-18]. https://arxiv.org/abs/2101.05783. DOI: 10.48550/arXiv.2101.05783.
[33] RAE J W, BORGEAUD S, CAI T, et al. Scaling language models: Methods, analysis & insights from training gopher[M/OL]. arXiv, 2022[2024-07-18]. https://arxiv.org/abs/2112.11446. DOI: 10.48550/arXiv.2112.11446.
[34] LI X, LI Y, JOTY S, et al. Evaluating psychological safety of large language models[M/OL]. arXiv, 2024[2024-07-18]. https://arxiv.org/abs/2212.10529. DOI: 10.48550/arXiv.2212.10529.
[35] JI Z, LEE N, FRIESKE R, et al. Survey of hallucination in natural language generation[J]. ACM Computing Surveys, 2023, 55(12): 1-38.
[36] HUANG S H, BHATIA K, ABBEEL P, et al. Establishing appropriate trust via critical states[M/OL]. arXiv, 2018[2024-07-18]. https://arxiv.org/abs/1810.08174. DOI: 10.48550/arXiv.1810.08174.
[37] AWAD E, DSOUZA S, KIM R, et al. The moral machine experiment[J]. Nature, 2018, 563(7729): 59-64.
[38] 王俊敏. 韦伯的理性“进步”及其意义问题[J]. 社会学研究, 2011, 25(2): 102-133.
[39] FORSTER E M. The machine stops[M]. London: Penguin, 2011.
[40] 谢小云, 左玉涵, 胡琼晶. 数字化时代的人力资源管理: 基于人与技术交互的视角[J]. 管理世界, 2021, 37(1): 200-216.
[41] BALASUBRAMANIAN N, YE Y, XU M. Substituting human decision-making with machine learning: Implications for organizational learning[J]. Academy of Management Review, 2022, 47(3): 448-465.
[42] BOSTROM N. Superintelligence: Paths, dangers, strategies[M]. Oxford: Oxford University Press, 2014.
[43] PANSERA M, FRESSOLI M. Innovation without growth: Frameworks for understanding technological change in a post-growth era[J]. Organization, 2021, 28(3): 380-404.
[44] HAIDT J. The righteous mind: Why good people are divided by politics and religion[M]. New York: Pantheon, 2012.
[45] ZHANG Z, ZHANG C, LIU N, et al. Heterogeneous value alignment evaluation for large language models[M/OL]. arXiv, 2024[2024-07-18]. https://arxiv.org/abs/2305.17147. DOI: 10.48550/arXiv.2305.17147.
[46] BETTIS R A, PRAHALAD C K. The dominant logic: Retrospective and extension[J]. Strategic Management Journal, 1995, 16(1): 5-14.

[47] SMITH B C. The promise of artificial intelligence：Reckoning and judgment[M]. Cambridge, MA：MIT Press, 2019.
[48] WHITTAKER M, CRAWFORD K, DOBBE R, et al. AI now report 2018[M]. New York：AI Now Institute at New York University New York, 2018.
[49] 王天思. 大数据中的因果关系及其哲学内涵[J]. 中国社会科学, 2016(5)：22-42.
[50] 肖峰. 人工智能与认识论的哲学互释：从认知分型到演进逻辑[J]. 中国社会科学, 2020(6)：49-71.
[51] LUGMAYR A, STOCKLEBEN B, SCHEIB C, et al. Cognitive big data：Survey and review on big data research and its implications. What is really "new" in big data?[J]. Journal of Knowledge Management, 2017, 21(1)：197-212.
[52] 董春雨. 从机器认识的不透明性看人工智能的本质及其限度[J]. 中国社会科学, 2023(5)：148-166.
[53] WEICK K E. Sensemaking in organizations[M]. Thousand Oaks, CA：Sage, 1995.

Research on Value Alignment Mechanism of Large Language Models Based on Formal and Substantive Rationality

Cheng Cong[1,2], Chen Jiachen[1], Yan Lulu[1]
(1. School of Management, Zhejiang University of Technology, Hangzhou 310023, China;
2. China Institute for SMEs, Zhejiang University of Technology, Hangzhou 310023, China)

Abstract: The value alignment of large language models is a global issue related to ensuring safe collaboration when enterprises and societies adopt these technologies. Achieving alignment between the behavior of large language models and the value intentions of decision-makers as well as societal norms is identified as the core challenge for ensuring safety and trust. Formal rationality and substantive rationality, two philosophical concepts proposed by Max Weber, were introduced to explore value alignment mechanisms. Four value alignment states in enterprise management were categorized including "high formal rationality-low substantive rationality" as technical drift, "high substantive rationality-low formal rationality" as value prioritization, "low formal rationality-low substantive rationality" as alignment failure, and "high formal rationality-high substantive rationality" as dynamic alignment. Transparency, clarity, and sociality were identified as analytical standards for value alignment. Pathways to achieve value alignment in enterprise management were proposed, including the embodiment of cognitive capability in the "technical drift→dynamic alignment" pathway, the clarification of technical intentionality in the "value prioritization→dynamic alignment" pathway, and the construction of meaning in the "alignment failure→dynamic alignment" pathway. The findings provide theoretical support and practical insights into the value alignment mechanisms of large language models in enterprise management.

Keywords: large language models technology; value alignment; formal rationality; substantive rationality